VOD 講義の話題区間の共起語による類似性検出

Detecting similarity for topic segment in VOD Lecture using Co-occurrence Word

中村 槙吾[*1], 椎名 広光[*2], 北川 文夫[*2], 小林 伸行[*3]
Shingo NAKAMURA[*1], Hiromitsu SHIINA[*2], Fumio KITAGAWA[*2], Nobuyuki KOBAYASHI[*3]
[*1] 岡山理科大学大学院 総合情報研究科 情報科学専攻
[*1]Graduate School of Informatics, Okayama University of Science
[*2] 岡山理科大学 総合情報学部 情報科学科
[*2]Faculty of Informatics, Okayama University of Science
[*3] 山陽学園大学 総合人間学部 生活心理学科
[*3]Faculty of Human Sciences, Sanyo Gakuen University
Email: 626653@teamgear.net

あらまし： VOD 講義の内容に対する検索機能として，我々は字幕データに対する検索語の出現頻度をもとに近似分布の当てはめを行い，その成分から利用者の意図する映像区間の推定を行っている．本研究では共起語に着目し，講義スライドのページ単位の共起語の出現確率からコサイン類似度による類似性を調査した．また、類似度の連続性から関連区間についても推定できるようにした．
キーワード：VOD 講義，コサイン類似度，共起度，区間推定

1. はじめに

現在，インターネット環境を利用して講義を行う VOD 講義が多くの大学で行われている．しかしながら，現状のシステムでは VOD の内容に対する検索機能がほとんど作成されていないため，講義のタイトルからいくつか候補を選び，動画を再生して目的のコンテンツを探す必要ある．これまでに我々は字幕データに対する検索語の出現頻度をもとにし，検索語が現れる確率に混合正規分布や混合ベータ分布に当てはめ，得られる近似分布の成分から利用者の意図する映像区間の推定を行っている．それらを調査する上で，検索語と検索語に共起する単語で推定区間の部分的に一致する関係が見られた．そこで本研究では，講義スライドのページ単位で字幕の単語の出現確率を求め，コサイン類似度を用いてスライドごとの類似性と，単語の出現確率に単語の共起度から求められる共起距離を反映させたスライド間の類似性を調査した．

2. VOD システムによる e-Learning 講義システム

本研究で作成しているシステムは，岡山理科大学を含む関連 6 大学で構成している教育コンソーシアム[1]における単位互換制度を利用した VOD による e-Learning 講義のシステム上に別途追加する形で開発している．

2007 年度データベースの講義では，1 回の講義は 3 つのセクションに分かれており，1 つのセクションは 20～30 分程度となっている．また，各セクションの最後に講義内容に関する課題があり，講義の内容の理解を確認するために用いられる．

3. 講義の話題の連続性の調査

講義の話題の関連性を調べるにあたり，スライドが切り替わるときに話題の変化がある．そこで，スライド単位の発話内容の類似性をコサイン類似度により調査した．

3.1 コサイン類似度

1 つのスライドにおける字幕の単語の出現確率をベクトルで表し，スライド x の字幕単語w_iの出現確率をx_i，スライド y の字幕単語w_iの出現確率をy_iとするとき，類似性$cos\theta$を次の式で表す．

$$cos\theta = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

全ての単語の出現確率からスライドごとのコサイン類似度を調査した結果を表 1 に示す．

3.2 共起度

映像区間の推定を行う上で検索語と検索語の共起語には少なからず関連性が見られる．そこで関連性の大きさを決めるために共起度を定義する．単語間の関連性の強さを表す共起度の算出法を以下に示す．

スライドの字幕全体 D，検索語w_i，検索語と共起する語w_j，w_iが出現する文S_i，w_iと 1 つのw_jの文節の差$D(w_i, w_j)$，w_jの頻度$frq(w_i)$とするとき，共起度$Cov(w_i, w_j)$を次の式で表す．

$$Cov(w_i, w_j) = \sum_{S_i \in D} \sum_{\omega_j \in S_i} \frac{\sqrt{frq(w_j)}}{D(w_i, w_j) + 1}$$

検索語を「キーワード」としたとき，上記の式で得られた共起度の高いもの 5 件（「広告」,「サイト」,「一つ」,「ビジネス」,「関連」）が選ばれる．スライ

表 1：単語の出現確率によるスライドの類似度

スライド番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	0.14	0	0	0	0.04	0	0	0.06	0	0.08	0	0	0	0.01	0.35	0.37
2	0.14	1	0.34	0.15	0.18	0.23	0.21	0.21	0.22	0.48	0.39	0.23	0.17	0.19	0.23	0.5	0.21
3	0	0.34	1	0.58	0.31	0.44	0.59	0.44	0.51	0.56	0.59	0.43	0.36	0.29	0.27	0.48	0.16
4	0	0.15	0.58	1	0.5	0.26	0.48	0.39	0.49	0.21	0.33	0.2	0.24	0.25	0.27	0.14	0.09
5	0	0.18	0.31	0.5	1	0.2	0.19	0.22	0.23	0.21	0.21	0.12	0.15	0.22	0.2	0.09	0.06
6	0.04	0.23	0.44	0.26	0.2	1	0.53	0.59	0.39	0.36	0.47	0.39	0.47	0.37	0.32	0.29	0.2
7	0	0.21	0.59	0.48	0.19	0.53	1	0.62	0.75	0.37	0.52	0.36	0.37	0.28	0.27	0.28	0.12
8	0	0.21	0.44	0.39	0.22	0.59	0.62	1	0.57	0.34	0.4	0.23	0.37	0.24	0.3	0.21	0.14
9	0.06	0.22	0.51	0.49	0.23	0.39	0.75	0.57	1	0.31	0.49	0.29	0.33	0.25	0.28	0.22	0.15
10	0	0.48	0.56	0.21	0.21	0.36	0.37	0.34	0.31	1	0.61	0.45	0.31	0.32	0.31	0.54	0.15
11	0.08	0.39	0.59	0.33	0.21	0.47	0.52	0.4	0.49	0.61	1	0.66	0.52	0.43	0.47	0.48	0.27
12	0	0.23	0.43	0.2	0.12	0.39	0.36	0.23	0.29	0.45	0.66	1	0.7	0.53	0.55	0.37	0.18
13	0	0.17	0.36	0.24	0.15	0.47	0.37	0.37	0.33	0.31	0.52	0.7	1	0.45	0.54	0.19	0.19
14	0	0.19	0.29	0.25	0.22	0.37	0.28	0.24	0.25	0.32	0.43	0.53	0.45	1	0.51	0.18	0.19
15	0.01	0.23	0.27	0.27	0.2	0.32	0.27	0.3	0.28	0.31	0.47	0.55	0.54	0.51	1	0.18	0.18
16	0.35	0.5	0.48	0.14	0.09	0.29	0.28	0.21	0.22	0.54	0.48	0.37	0.19	0.18	0.18	1	0.2
17	0.37	0.21	0.16	0.09	0.06	0.2	0.12	0.14	0.15	0.15	0.27	0.18	0.19	0.19	0.18	0.2	1

表 2：単語の出現確率(検索語：キーワード)

	検索語	共起語1	共起語2	共起語3	共起語4	共起語5
スライド番号	キーワード	広告	サイト	一つ	ビジネス	関連
1	0	0	0	0	0	0
2	0	0	0	0	1	0
3	0	0	0.833333	0.166667	0	0
4	0	0	1	0	0	0
5	0	0	1	0	0	0
6	0.857143	0	0.142857	0	0	0
7	0.1	0	0.9	0	0	0
8	0	0	1	0	0	0
9	0.076923	0	0.923077	0	0	0
10	0	0	0	0	0	0
11	0.333333	0.333333	0.25	0.083333	0	0
12	0.444444	0.388889	0.111111	0	0	0.055556
13	0.266667	0.6	0.133333	0	0	0
14	0.5	0.3	0.1	0.1	0	0
15	0.25	0.6	0.1	0	0	0.05
16	0	0	0	0	0.666667	0.333333
17	0.25	0.25	0	0.25	0.25	0

ドごとに，「キーワード」とそれらの共起語の出現確率を表 2 に示す．

3.3 共起距離による重み付け

表 2 に示している結果には，検索語と共起語の類似度である共起度について考慮されていない．そこで共起度から共起距離を定義し，それによって重み付けを行った．共起距離の定義は次の通り．

$$共起距離 = \frac{検索語の共起度}{共起語の共起度} = \frac{Cov(w_i, w_i)}{Cov(w_i, w_j)}$$

講義のセクション全体で検索語とその共起語の頻度を調べ，スライドごとに出現確率を求める．次にスライドの単語の出現確率に共起距離による重み与え，ベクトルとして，コサイン類似度により関連性を求める．共起距離を乗じたスライドごとの単語の出現確率を表 3，表 3 から導いたコサイン類似度を表 4 に示す．

4. 類似の比較

連続してスライドの類似度が高くなっている範囲について，表 1 と表 4 で比較を行った．まず表 1 ではスライド 6～8, 10～15 のあたりで多少の類似関係が見られる．それに対して表 4 では関係が大きいであろうスライドの類似度が 0.7 以上の高い数値となり，より特徴的な結果となっている．特にスライド 3～5，7～9，11～15 はそれぞれの範囲のスライド全ての類似度が高くなっている．また，スライド 6 は離れているがスライド 11～15 にかけて高い類似度が出ている．これらの結果について，実際に VOD

表 3：共起距離による重み付けを行った単語の出現確率(検索語：キーワード)

	検索語	共起語1	共起語2	共起語3	共起語4	共起語5
スライド番号	キーワード	広告	サイト	一つ	ビジネス	関連
1	0	0	0	0	0	0
2	0	0	0	0	1	0
3	0	0	0.855909	0.144091	0	0
4	0	0	1	0	0	0
5	0	0	1	0	0	0
6	0.974766	0	0.025234	0	0	0
7	0.417026	0	0.582974	0	0	0
8	0	0	1	0	0	0
9	0.349173	0	0.650827	0	0	0
10	0	0	0	0	0	0
11	0.678223	0.2206	0.079009	0.022168	0	0
12	0.747614	0.212774	0.029031	0	0	0.010581
13	0.552639	0.404442	0.04292	0	0	0
14	0.798486	0.15583	0.024805	0.020879	0	0
15	0.536077	0.418476	0.033307	0	0	0.01214
16	0	0	0	0	0.679623	0.320377
17	0.634477	0.206371	0	0.082954	0.076198	0

表 4：共起距離による重み付けを行った場合の類似度(検索語：キーワード)

スライド番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1																	
2	1	0	0	0	0	0	0	0	0		0	0	0	0	0	0.9	0.11
3	0	1	0.99	0.99	0.03	0.8	0.99	0.87			0.11	0.04	0.06	0.03	0.05	0	0.02
4	0	0.99	1	1	0.03	0.81	1	0.88			0.11	0.04	0.06	0.03	0.05	0	0
5	0	0.99	1	1	0.03	0.81	1	0.88			0.11	0.04	0.06	0.03	0.05	0	0
6	0	0.03	0.03	0.03	1	0.6	0.03	0.5			0.95	0.96	0.81	0.98	0.79	0	0.94
7	0	0.8	0.81	0.81	0.6	1	0.81	0.99			0.64	0.59	0.52	0.6	0.5	0	0.55
8	0	0.99	1	1	0.03	0.81	1	0.88			0.11	0.04	0.06	0.03	0.05	0	0
9	0	0.87	0.88	0.88	0.5	0.99	0.88	1			0.54	0.49	0.44	0.49	0.42	0	0.44
10																	
11	0	0.11	0.11	0.11	0.95	0.64	0.11	0.54			1	1	0.95	0.99	0.94	0	0.98
12	0	0.04	0.04	0.04	0.96	0.59	0.04	0.49			1	1	0.94	1	0.93	0.01	0.98
13	0	0.06	0.06	0.06	0.81	0.52	0.06	0.44			0.95	0.94	1	0.9	1	0	0.93
14	0	0.03	0.03	0.03	0.98	0.6	0.03	0.49			0.99	1	0.9	1	0.89	0	0.98
15	0	0.05	0.05	0.05	0.79	0.5	0.05	0.42			0.94	0.93	1	0.89	1	0.01	0.93
16	0.9	0	0	0	0	0	0	0			0	0.01	0	0	0.01	1	0.1
17	0.11	0.02	0	0	0.94	0.55	0	0.44			0.98	0.98	0.93	0.98	0.93	0.1	1

を見て調査した．

VOD を見た結果，スライド 3～5，7～9，11～15 がそれぞれ一つの話題，スライド 6 はそれ以降の話題全体の概要であると分かった．ただし，スライド 3，7 については次の話題の概要でもあるとも言える．

以上のことから共起距離による重み付けを行った場合には，特に関係の強い区間が得られると考えられる．

5. まとめ

本研究では VOD 講義のスライドの区間で，コサイン類似度により話題の類似性を調査した．また検索語を決め，検索語とその共起語の出現確率に共起距離による重み付けを行った場合も調査した．その結果として，重み付けを行った調査では特に似ている区間を得られた．ただし，今回の結果は一つの検索語だけで調査したものであり，他の単語では違う結果が出る可能性がある．今後の課題として多くの検索語，また他のセクションでも試してみたい．

参考文献

(1) 北川，大西: “対面講義と e-learning(LMS + VOD) とを併用した講義形式の実践と分析”，日本教育情報学会学会誌 Vol.22 No.3, pp.57-66（2007）
(2) 伊藤，藤井，石川: “音声文書検索を用いたオンデマンド講義システム”，電子情報通信学会技術研究報告 SP 音声,Vol.101, No.523, pp.55-60（2001）